Optimizing Large Data Transfers over 100Gbps Wide Area Networks

Anupam Rajendran, Parag Mhashilkar, Hyunwoo Kim, Dave Dykstra, Gabriele Garzoglio, Ioan Raicu

High Throughput Data Program

Fermi National Accelerator Laboratory

Ioan Raicu

Slides Prepared by Anupam Rajendran

May 14th, 2013 – IEEE CCGrid 2013

ILLINOIS INSTITUTE
OF TECHNOLOGY
Transforming Lives.Inventing the Future.www.iit.edu

Fermilab

Introductions

- Collaborative effort between Illinois Institute of Technology (IIT) and Fermi National Accelerator Laboratory (FNAL)
  - FNAL is lead institution
  - Large part of results obtained during Anupam Rajendran's (first author) summer 2012 internship while on site at Fermi
  - Anupam was supposed to present paper, but his travel VISA did not arrive in time
- Ioan Raicu @ IIT
  - Director of Data-Intensive Distributed Systems Laboratory @ IIT

Fermi National Accelerator Laboratory

- Fermi National Accelerator Laboratory Mission
  - Advances the understanding of the fundamental nature of matter and energy by providing leadership and resources for qualified researchers to conduct basic research at the frontiers of high energy physics and related disciplines

Fermilab Research Activities

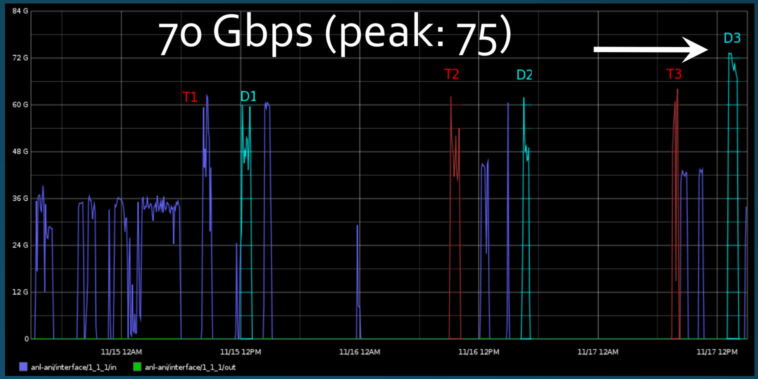
- Fermilab hosts the US Tier-1 Center for the LHC's(Large Hadron Collider) Compact Muon Solenoid (CMS) experiment – Store, process and again distribute data
- Using the network for decades in the process of scientific discovery for sustained, high speed, large and wide-scale distribution of and access to data
    - High Energy Physics community
    - Multi-disciplinary communities using grids (OSG, XSEDE)
- 94 Petabytes written to tape, today mostly coming from offsite
- 160Gbps peak LAN traffic from archive to local processing farms
- LHC peak WAN usage in/out of Fermilab at 20-30 Gbps
- Challenges in scaling and distribution of big data

High Throughput Data Program

- Experiment analysis systems include a deep stack of software layers and services
- Need to ensure these are functional and effective at the 100G scale end-to-end
  - Measure and determine efficiency of the end-to-end solutions
  - Determine and tune the configuration of all layers to ensure full throughput in and across each layer/service
  - Monitor, identify and mitigate error conditions

High Throughput Data Program

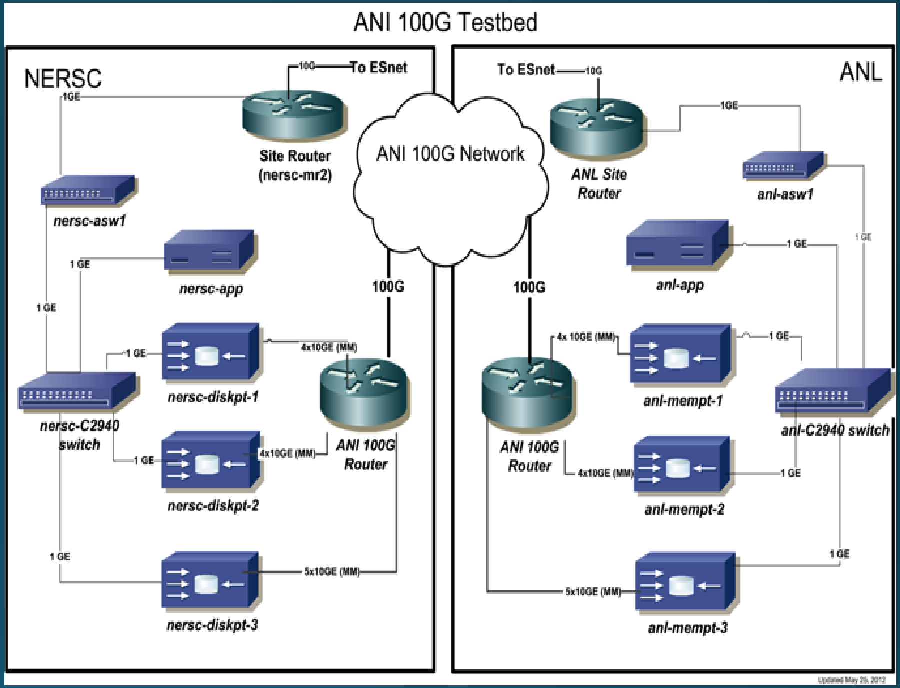
- 2011-2012
  - Advanced Network Initiative (ANI) Long Island MAN (LIMAN) testbed.
  - GridFTP and Globus Online tests over 3x10GE.
  - Super Computing '11
    - Demonstrated transfer of ~30TB of CMS data in 1h from NERSC to ANL using GridFTP on a shared 100G network



- 2012-2013: ESnet 100G testbed
  - Custom boot images
  - Tuning parameters of middleware for data movement: xrootd, GridFTP, SRM, Globus Online, Squid. Achieved ~97Gbps
- Summer 2013: 100G Endpoint at Fermilab
  - Validate hardware link w/ transfer apps for CMS current datasets
  - Test NFS v4 over 100G using dCache (collab. w/ IBM research)

Advanced Networking Initiative (ANI) Testbed

- Cross-country 100Gbps testbed linking DOE supercomputers in Argonne National Lab (ANL) and National Energy Research Scientific Computing Center (NERSC)



Test Hardware and Network

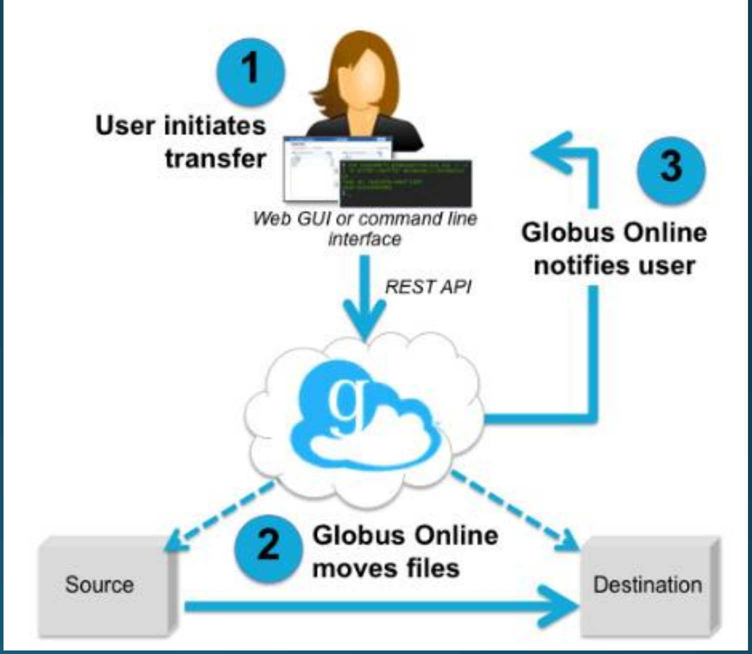
- ANL – AMD(2.6GHz) - 16 cores - 64GB
- NERSC 1, 2 – Intel (2.67 GHz) - 12 cores - 48 GB, NERSC 3 – Intel (2.4 GHz) - 8 cores - 24 GB
- 3 nodes per site
  - 4x10Gb/s network links
  - 120Gb/s internal bandwidth
  - 100Gb/s external bandwidth
- RTT between NERSC and ANL is measured to be 53 ms
- This testbed is not directly accessible via internet. For outside users it is available through Virtual Private Network using a Virtual Machine at Fermilab.

GridFTP

- A high performance, secure, reliable data transfer protocol optimized for high bandwidth, wide-area networks
- Provides a uniform way of accessing the data
- FTP was chosen because of its widespread use and was easier to add extensions
- **Globus Toolkit**: a reference implementation of GridFTP provides server, client tools and deployment libraries
- Features:
  - GSI Security for authentication and encryption to file transfers
  - Data channel reuse
  - Third-party transfers: C can initiate a transfer from A to B.
  - Parallel, Concurrent, Striped transfers
  - Partial file transfer, Restart failed transfers
  - Tunable network parameters

Globus Online

- Move, sync, share files
  - Easy “fire-and-forget” transfers
  - Share with any Globus user or group
  - Automatic fault recovery & high performance
  - Across multiple security domains
  - Web, command line, and REST interfaces
- • Minimize IT costs
  - Software as a Service (SaaS)
    - No client software installation
    - New features automatically available
  - Consolidated support & troubleshooting
  - Simple endpoint installation with Globus Connect and GridFTP

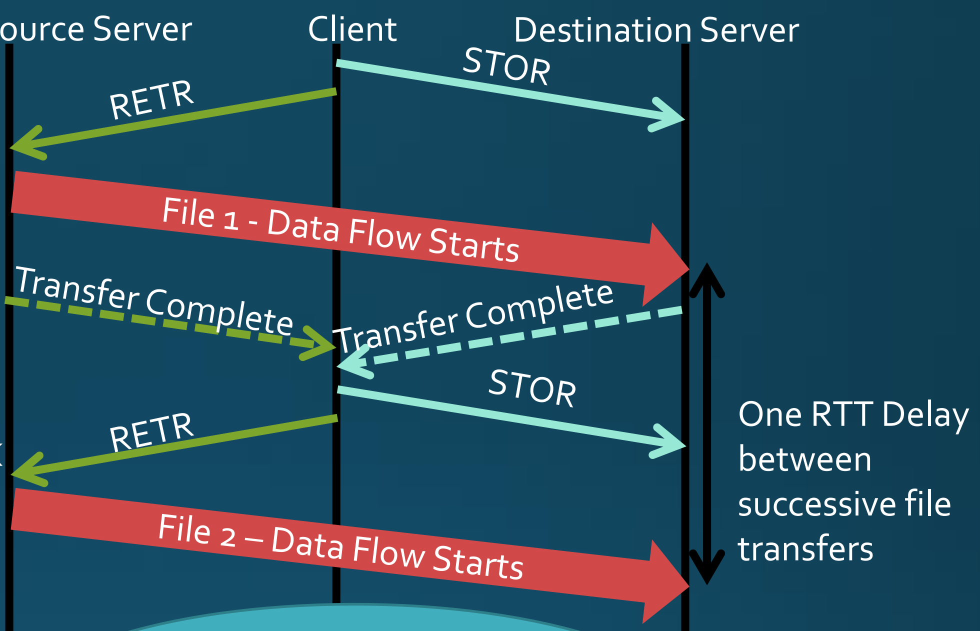

1
User initiates transfer
Web GUI or command line interface
REST API
3
Globus Online notifies user
2
Globus Online moves files
Source
Destination

Storage Resource Management(SRM)

- Common protocol for interfacing storage used for
  - Metadata operations
  - Data movement between storage elements
  - Generic management of backend storage
- Uses GridFTP for data transfer
- Effectively load balance transfers over multiple nodes and thus showing good scalability
- **BeStMan (server) and LCG Utilities (client)** were used for the testing

Analysis of File Transfer Operations

- GridFTP
  - Overhead of increased security and reliability
  - Lots Of Small Files(LOSF) problem
    - **GridFTP pipelining** does not work for list of files & supports directory transfers only
    - **Logging to disk** through 1Gbps link
- Globus Online
  - High control channel latency **~150ms**
- SRM
  - Performance bounded by GridFTP
  - Overhead of converting urls srm:// to gsiftp://
  - No data channel caching
  - No access to all the tuning parameters



Source Server

Client

Destination Server

STOR

RETR

File 1 - Data Flow Starts

Transfer Complete

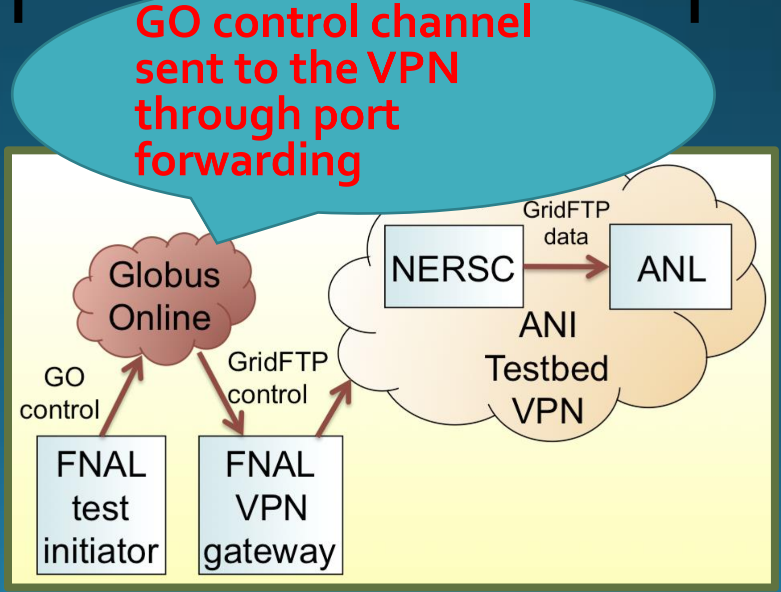
Transfer Complete

STOR

RETR

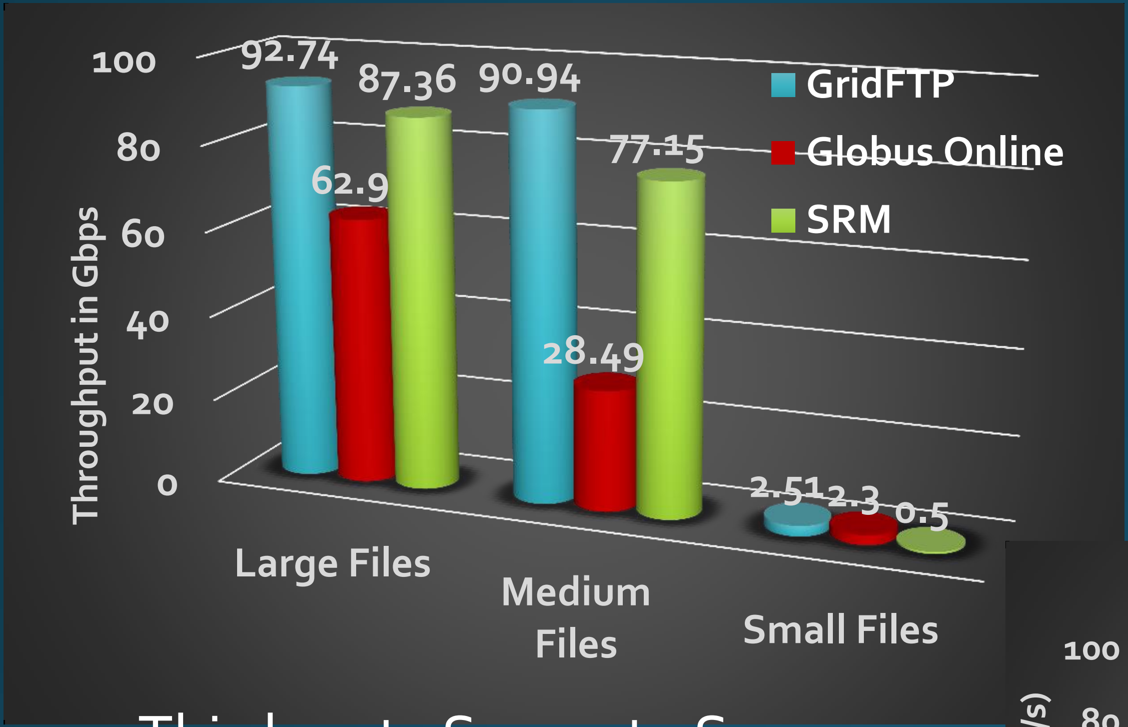
File 2 – Data Flow Starts

One RTT Delay between successive file transfers



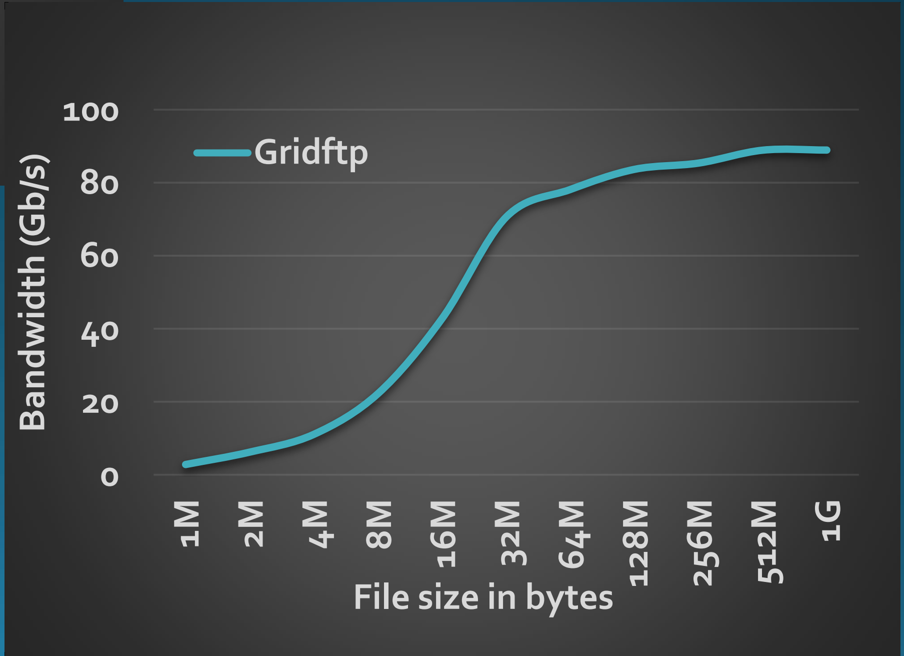
GO control channel sent to the VPN through port forwarding

GridFTP, Globus Online and SRM Performance on ANI Testbed



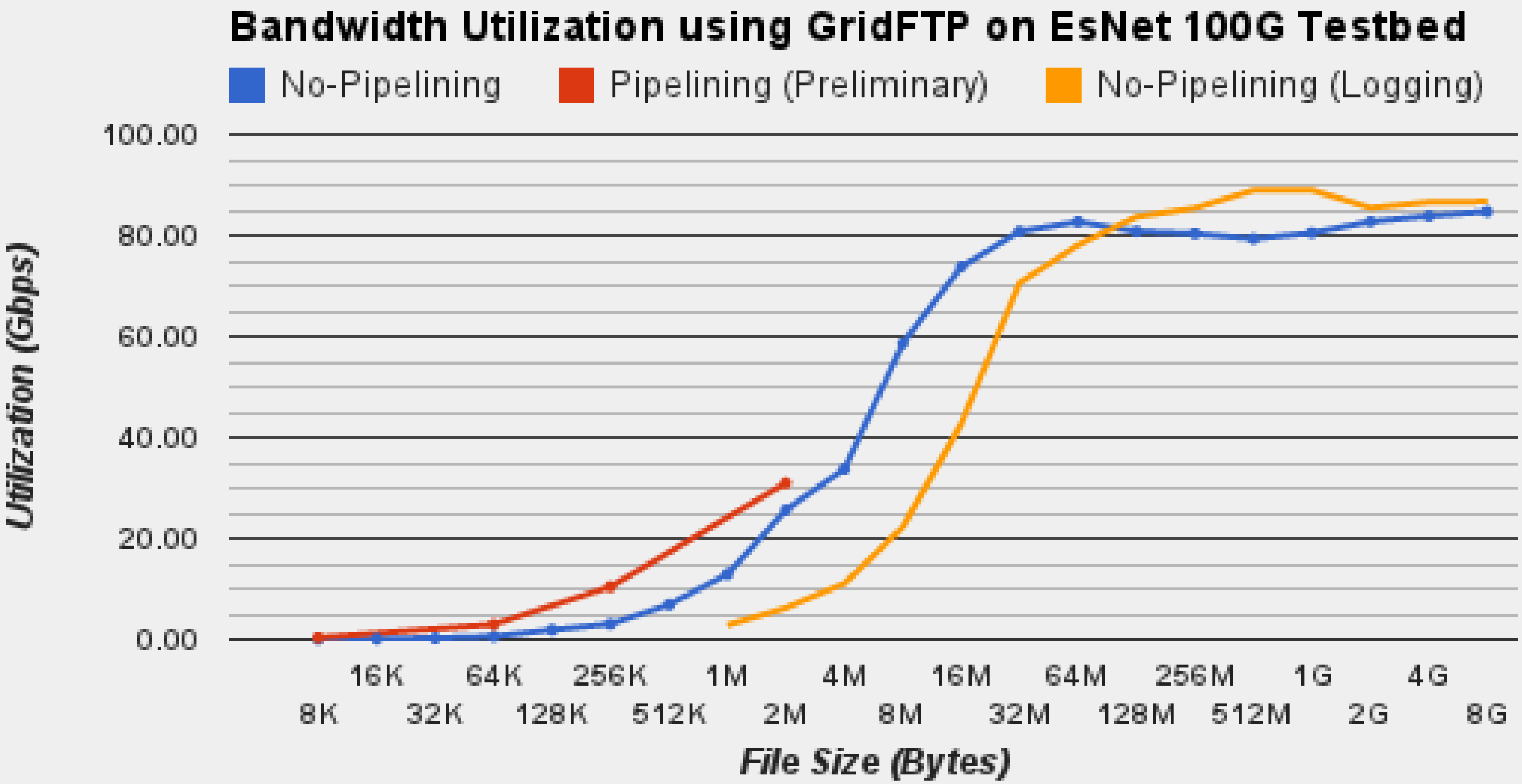
Data set

- Small - 8KB to 4MB
- Medium - 8MB to 1GB
- Large - 2GB to 8GB



- Third party Server to Server transfers:
  src at NERSC / dest at ANL
- Dataset split into 3 size sets
- Large files transfer performance **~ 92Gbps**

GridFTP Updated Results (not in paper)

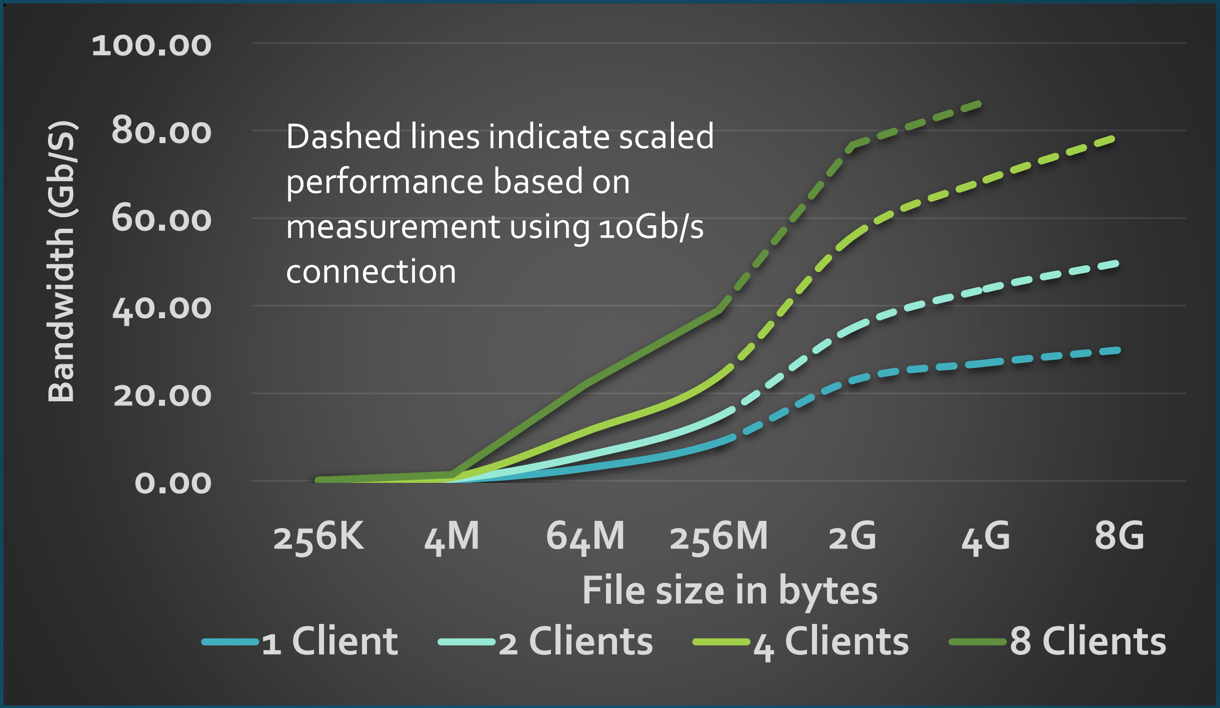


XrootD

- A file access and data transfer *protocol*
    - Defines POSIX-style byte-level random access for
        - *Arbitrary* data organized as files of *any* type
        - Identified by a hierarchical directory-like name
- It is not a POSIX file system – it does not provide full POSIX file system semantics
    - There is a FUSE implementation called xrootdFS
        - An xrootd *client* simulating a mountable file system
- It is not an Storage Resource Manager (SRM)
    - Provides SRM functionality via BeStMan

XrootD Performance on ANI 100Gbps Testbed

- Data Movement over XRootD, testing LHC experiment (CMS / Atlas) analysis use cases.
  - Clients at NERSC / Servers at ANL
  - Using RAMDisk as storage area on the server side
- Challenges
  - Tests limited by the size of RAMDisk
  - Little control over **xrootd** client / server tuning parameters



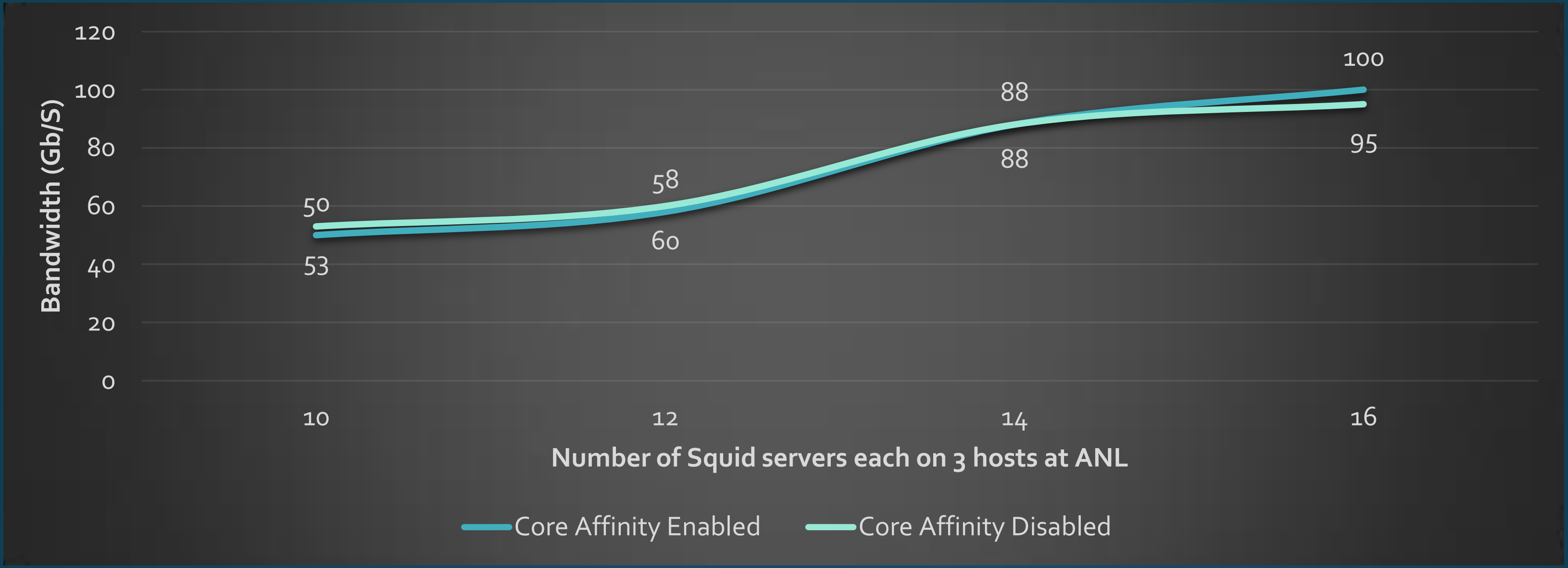
Dashed lines indicate scaled performance based on measurement using 10Gb/s connection

Dataset (GB)	1 NIC measurements (Gb/s)	Aggregate Measurements (12 NIC) (Gb/s)	Scale Factor per NIC	Aggregate estimate (12 NIC) (Gb/s)
0.512	4.5	46.9	0.87	-
1	6.2	62.4	0.83	-
4	8.7 (8 clients)	-	0.83	86.7
8	7.9 (4 clients)	-	0.83	78.7

Squid

- **Frontier Squid** – Proxy server/Web Cache
- Any Internet Object – file, document, HTTP/FTP response is cached
- Reduces bandwidth usage and improves response time through caching and reusing
- Load balance the web servers
- LRU algorithm replace items in cache
- Single threaded

Squid Performance on ANI 100Gb/s Testbed



- Used **wget(client)** to fetch 8MB file repeatedly from Squid server. This size mimics LHC use case for large calib. data ➔ 9000 clients
- Varying number of Squid instances running in each host
- both directions – NERSC ↔ ANL
- With and without Core Affinity

- **Core-affinity improves performance by 21% in some tests**
- Increasing the number of squid processes improves performance
- Best performance with 16 Squid servers and 9000 clients: **~100 Gbps**

Summary

- This work has been done in context of a broader FNAL work
  - Spans all layers of the communication stack for identifying the gaps in middleware used by HEP community
- Evaluated grid middleware on ANI 100G Testbed
  - Results indicates the potential of middleware technologies to scale up to 100Gbps
  - GridFTP – large files: 4 parallel TCP streams & 4 concurrent connections; small files: no pipelining for individual files, logging through slow n/w (although it could be disabled)
  - Globus Online – high control channel latency
  - SRM (small files) – no data channel caching
  - XrootD – Little control over xrootd client / server tuning parameters
  - Squid – 16 Squid servers with 9000 clients
- Fermilab should have production 100GE capability by Summer 2013